ADLHW3 Report

Student ID: R12922184 Name:鄭星逸

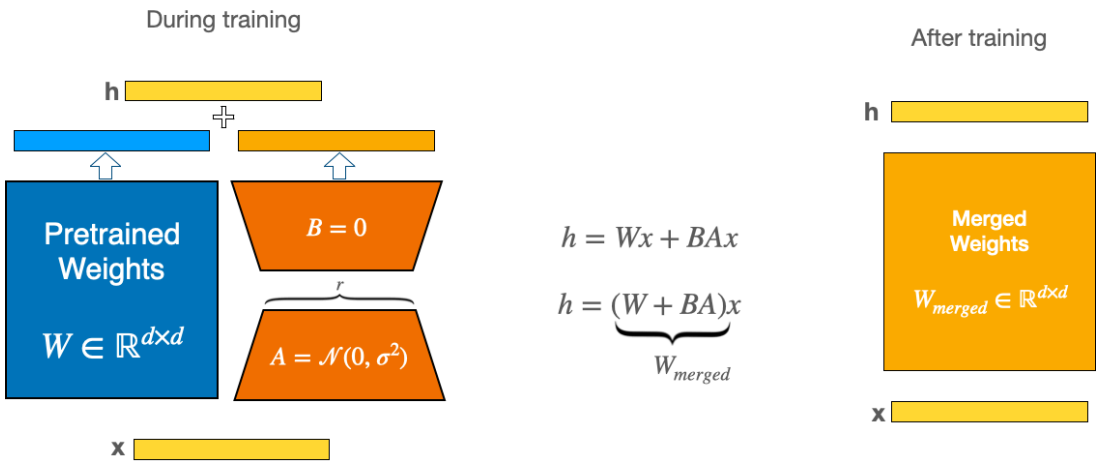
Q1: LLM Tuning

How much training data did you use? (2%)

Answer:
All 10000 training data in train.json.

How did you tune your model? (2%)

Answer:
I referred to the qlora.py on GitHub, which is at https://github.com/artidoro/qlora.
I fine-tuned the model using LoRA (Low-Rank Adaptation) by integrating LoRA modules into the key, query, value, and output projection layers (k_proj, q_proj, v_proj, o_proj). To optimize efficiency, we applied 4-bit quantization using the bitsandbytes library, reducing model size and speeding up training.
I used the Hugging Face Transformers library to handle model loading and tokenization. The data preprocessing involved constructing input-output pairs suitable for causal language modeling, ensuring that the inputs and targets were appropriately tokenized and padded.



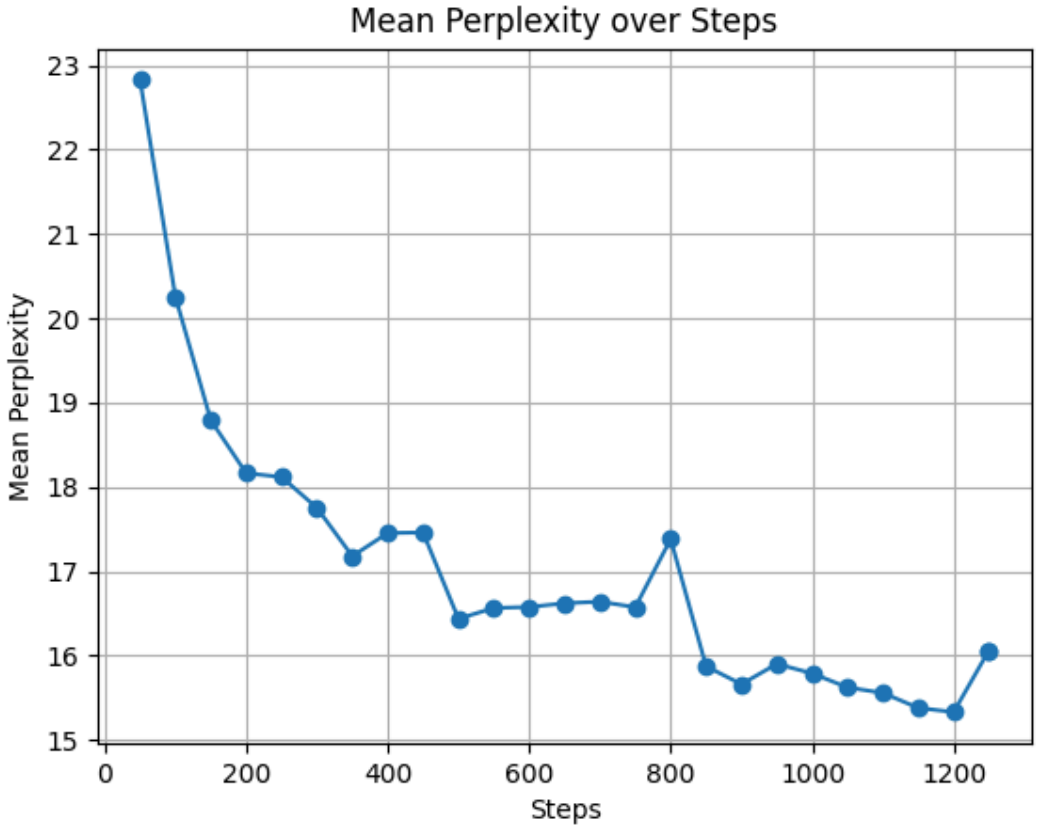
source from:
https://www.linkedin.com/pulse/lora-low-rank-adaptation-arjun-p-v-ljn4c/

What hyper-parameters did you use? (2%)
Answer:
The hyperparameters are shown in the table below, and I use the default prompt.

model name	"zake7749/gemma-2-2b-it-chinese-kyara-dpo"
batch size	4
gradient accumulation steps	4
max steps	1250
learning rate	0.0001
fp16	true
bits	4
lora_r	32
lora_alpha	16
lora_dropout	0.1
warmup ratio	0.03
warmup steps	50

What is the final performance of your model on the public testing set? (2%)
Answer:
Mean perplexity: 16.468677958488463

Plot the learning curve on the public testing set (2%)
Answer:
I use ppl.py as the evaluation function to record the mean perplexity every 50 steps. The curve is shown in below.



Q2: LLM Inference Strategies

- *Zero-Shot*

  What is your setting? How did you design your prompt? (1%)

Answer:
I utilized ppl.py as the evaluation script with the default prompt settings.
"你是人工智慧助理，以下是用戶和人工智能助理之間的對話。你要對用戶的問題提供有用、安全、詳細和禮貌的回答。USER: {instruction} ASSISTANT:"
Mean perplexity: 10354.311291719436

Subsequently, I modified the prompt as follows:
"你是一個文言文翻譯專家，以下為用戶和文言文翻譯專家的對話，請提供準確清晰完整，有邏輯的回答。USER: {instruction} ASSISTANT:"
Mean perplexity: 4744.97582774353

- *Few-Shot (In-context Learning)*

  What is your setting? How did you design your prompt? (1%)

  How many in-context examples are utilized? How you select them? (1%)

Answer:
I utilized ppl.py as the evaluation script with the modified prompt as follows.

Firstly, I utilized one in-context example by randomly selecting a data pair from the training dataset as the example.

"你是一個文言文翻譯專家, 以下為用戶和文言文翻譯專家的對話, 請提供準確清晰完整, 有邏輯的回答。以下是一個例子，"instruction": "高祖初, 為內秘書侍禦中散。\n翻譯成現代文:", output": "高祖初年, 任內秘書侍禦中散。USER: {instruction} ASSISTANT:"
Mean perplexity: 8190.474561386109

Secondly, I utilized two in-context examples: one for translating from Classical Chinese (文言文) to Vernacular Chinese (白话文), and the other for translating from Vernacular Chinese to Classical Chinese. The prompt is shown as follows.

"你是一個文言文翻譯專家, 以下為用戶和文言文翻譯專家的對話, 請提供準確清晰完整, 有邏輯的回答。以下是兩個例子, 1. instruction: 高祖初, 為內秘書侍禦中散。\n翻譯成現代文:, output: 高祖初年, 任內秘書侍禦中散。2. instruction:它的旁邊有一顆小星, 名叫長沙星, 星不宜明, 若與軫宿的四顆星一樣明亮, 五顆星進入軫宿, 錶示將有大的戰爭發生。\n這句話在中國古代怎麼說:output: 其旁有一小星, 曰長沙, 星星不欲明;明與四星等, 若五星入軫中, 兵大起。USER: {instruction} ASSISTANT:"
Mean perplexity: 4031.0294188671114

- *Comparison:*

  What's the difference between the results of zero-shot, few-shot, and LoRA? (2%)

Answer:

prompt	ppl Mean perplexity
zero-shot(default)	10354.31
zero-shot(improved)	4744.97
few-shot(one example)	8190.47
few-shot(two examples)	4031.02
LoRA	16.46

This table presents the mean perplexity scores for different prompting strategies. The "zero-shot (default)" approach has a mean perplexity of 10354.31, while the "zero-shot (improved)" method shows a reduced perplexity of 4744.97. The "few-shot (one example)" strategy results in a perplexity of 8190.47, and the "few-shot (two examples)" approach achieves the lowest perplexity of 4031.02. These results indicate that using improved zero-shot and few-shot examples can significantly decrease perplexity, enhancing model performance.

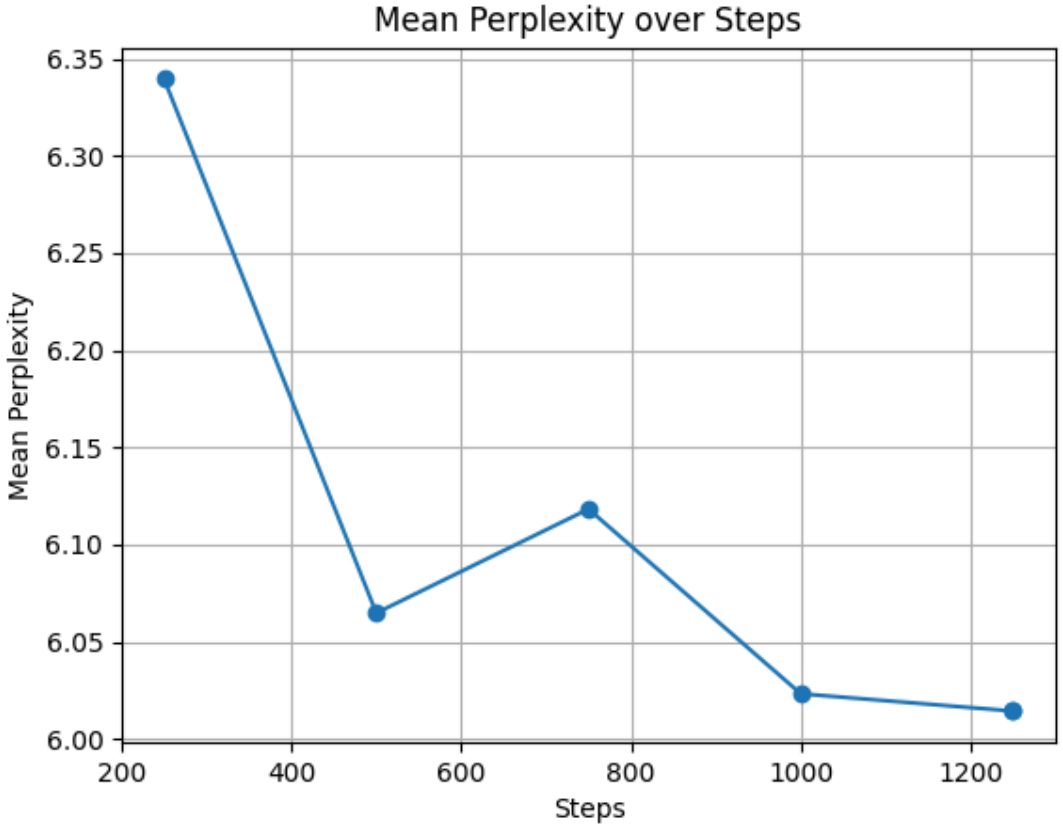
Q3: Bonus: Try Llama3-Taiwan (8B) (2%)

Llama-3-8b trained by traditional Chinese data

Tune this model on the classical chinese data

Describe your experimental settings and compare the results to those obtained from your original methods

Answer:
I use the same hyperparameters as shown in Q1, but the model name is "yentinglin/Llama-3-Taiwan-8B-Instruct." And evaluate model every 250 steps. The curve is shown in below.



The final score is Mean perplexity: 6.239566230773926 which is much better than gemma.